UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 13-29
Statistics and Econometrics Series 25
Septiembre 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

MODELLING LONG TERM TREND AND LOCAL SPATIAL CORRELATION: A MIXED PENALIZED SPLINE AND SPATIAL ECONOMETRICS APPROACH

Román Mínguez[1],María Durbán[2], José María Montero[3],
Dae-Jin Lee[4]

Abstract:

In this work we propose the combination of P-splines with traditional spatial econometric models in such a way that it allows for their representation as a mixed model. The advantages of combining these models include: (i) dealing with complex non-linear and non-separable trends, (ii) estimating short-range spatial correlation together with the large-scale spatial trend, (iii) decomposing the systematic spatial variation into those two components and (iv) estimating the smoothing parameters included in the penalized splines together with the other parameters of the model. The performance of the proposed spatial non-parametric models is checked by both simulation and a empirical study. More specifically, we simulate 3,600 datasets generated by those models (with both linear and non-linear-non-separable global spatial trends). As for the empirical case, we use the well-known Lucas county data on housing prices. Our results indicate that the proposed models have a better performance than the traditional spatial strategies, specially in the presence of nonlinear trend.

Keywords: global spatial trend, mixed models, P-splines, PS-SAR, PS-SEM.

[1] ,[3] Department of Statistics, Universidad Castilla-La Mancha, Toledo, Spain

[2] Department of Statistics, Universidad Carlos III de Madrid, Spain

[4] CSIRO Mathematics, Informatics and Statistics, Clayton, VIC, Australia

Modelling long term trend and local spatial correlation: a mixed penalized spline and spatial econometrics approach.

Abstract

In this article we propose the combination of P-splines with traditional spatial econometric models in such a way that it allows for their representation as a mixed model. The advantages of combining these models include: (i) dealing with complex non-linear and non-separable trends, (ii) estimating short-range spatial correlation together with the large-scale spatial trend, (iii) decomposing the systematic spatial variation into those two components and (iv) estimating the smoothing parameters included in the penalized splines together with the other parameters of the model. The performance of the proposed spatial non-parametric models is checked by both simulation and a empirical study. More specifically, we simulate 3,600 datasets generated by those models (with both linear and non-linear-non-separable global spatial trends). As for the empirical case, we use the well-known Lucas county data on housing prices. Our results indicate that the proposed models have a better performance than the traditional spatial strategies, specially in the pesence of non-linear trend.

Keywords: global spatial trend, mixed models, P-splines, PS-SAR, PS-SEM.

JEL classification: C14, C15, C21.

AMS classification: 91B72, 93E14, 65D07, 62M30, 62G08.

1. Introduction

We agree with Cressie and Wikle (2011) that questions related with Climate and Environment and those related with Education, Health and Economics are fundamental to sustaining our planet (the first ones) and improve social and economic conditions (the last ones). Since these questions are the Society's dominant questions in the current century, this century should be a century of massive (spatial and spatio-temporal) datasets collected to answer such questions, which are inherently statistical and econometrics. Thus, the analysis of spatial (and more recently spatio-temporal) data is currently of great interest to statistical modeling, especially from the econometrics and geostatistics perspectives.

Here we focus on the econometrics perspective, which usually demands a number of explanatory variables large enough to yield good predictions. However, despite the above, until now, unfortunately, when dealing with real databases, the number of explanatory variables is not as large as desired, and as a consequence the results are not so good as expected. But in case that there is spatial heterogeneity and/or spatial dependence in the data, the problem of a reduced number of explanatory variables can be solved by taking advantage of the estimation of such heterogeneity (spatial trend) and spatial dependence (spatial dependence parameter), for which only the spatial coordinates are needed.

Since this is the case in a large number of real situations, in this article we propose a new family of spatial semiparametric models that accounts for both heterogeneity —by incorporating a complex non-linear trend (non specified a priori)— and spatial dependence —by adding to the trend and covariates a Spatial Autoregressive (SAR) model or a Spatial Error Model (SEM). The advantage of this class of models is that they account for both the short and large scale variation, and, most importantly, they are capable of perfectly distinguish

both types of variation.

In brief, our extended proposal is the combination of a nonparametric trend (not specified a priori, unlike parametric trends) with a SAR or SEM model in a novel specification where all parameters (included the smoothing parameter associated to the nonparametric trend) are jointly estimated by REML in a single stage. That is, PS-spatial models jointly account for spatial heterogeneity and spatial dependence, being able of distinguish both spatial characteristics. This is important per se, but also because in case of a reduced number of covariates predictions are really good.

This is precisely the case of house price prediction. Although house price prediction is currently a core task in many countries, unfortunately house price databases only include a few traditional covariates. This is why in this paper we put PS-SAR and PS-SEM models at work competing with other traditional strategies, to show that in case of a reduced number of covariates in a spatial context these novel strategies are really attractive in light of their good predictions.

The paper is organized as follows. Section 2 introduces P-splines and their mixed model representation in one and two dimensions. Section 3 extends the traditional spatial econometric specifications by incorporating a two-dimensional P-spline to account for the large-scale trend. A simulation study is carried out in section 4 in order to compare the performance of the spatial econometric models with the new specifications proposed. Section 5 includes an empirical case study using the well-known Lucas county (Ohio) housing price database, and we conclude with a discussion in section 6.

2. Spatial smoothing with P-splines

2.1. Introduction to P-splines

For the sake of clarity, we first introduce P-splines in the one dimensional setting, before extending them to spatial smoothing.

Suppose the variable $\mathbf{y}$ depends smoothly on a covariate $\mathbf{x}$, then, a smooth model for the data would be given by:

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1}$$

and most smooth models differ on how $f(\mathbf{x})$ is estimated. Here we take the Eilers and Marx (1996) approach, that is: i) $f(\mathbf{x}) = \mathbf{B}\boldsymbol{\theta}$, where $\mathbf{B} = (B_1(\mathbf{x}), B_2(\mathbf{x}), \ldots, B_k(\mathbf{x}))$ is an $n \times k$ regression matrix of B-splines (see De Boor, 1977), and $\boldsymbol{\theta}$ is the corresponding vector of coefficients; ii) the coefficients of adjacent columns in the basis satisfy certain smoothness conditions which can be expressed in terms of finite differences of the θ_i's. The penalty matrix is based on the differences of adjacent coefficients $\mathbf{P} = \lambda\mathbf{D}'\mathbf{D}$, where $\mathbf{D}$ is a difference matrix (in general, differences of order two are used), and λ is a smoothing parameter. Then, the coefficients $\boldsymbol{\theta}$ are chosen to minimize:

$$S(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}.$$

Minimizing S leads to

$$\hat{\boldsymbol{\theta}} = (\mathbf{B}'\mathbf{B} + \mathbf{P})\mathbf{B}'\mathbf{y},$$

As can be seen, the solution of the set of equations above depends on λ. When it comes to model estimation, previous approaches estimated λ separately using cross-validation, generalized cross-validation etc. procedures. However, we take a different approach based on the mixed model representation of penalized splines that allows the join estimation of λ and the rest of model parameters. More specifically, given model (1), and based on the singular value decomposition of the penalty matrix $\mathbf{D}'\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}'$, we define:

$$\mathbf{X} = [1 : \mathbf{x}]; \quad \mathbf{Z} = \mathbf{B}\mathbf{U}_s\boldsymbol{\Sigma}_s^{-1/2}. \tag{2}$$

Then, following Currie and Durbán (2002) model (1) can be immediately rep-

resented as a mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \quad \mathbf{u} \sim N(0, \sigma_\alpha^2 \mathbf{I}_{c-2}) \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}), \tag{3}$$

where the smoothing parameter appears implicitely as $\lambda = \sigma^2/\sigma_\alpha^2$, and can be estimated jointly with the other parameters of the model using the REML approach as usual in standard mixed models methodology.

2.2. P-splines for spatial data

In the case of spatial data $(\mathbf{x}_{1i}, \mathbf{x}_{2j}, \mathbf{y}_{ij})$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ indicate the geographic longitude and latitude, respectively, and $\mathbf{y}$ is the response variable, model (1) turns into:

$$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2) + \boldsymbol{\epsilon} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{4}$$

where now $\mathbf{B}$ is a matrix of regression basis constructed from the covariates $(\mathbf{x}_1, \mathbf{x}_2)$.

As a consequence, it follows three extensions of the univariate model:

1. The basis is constructed from the tensor product of B-spline basis with equally spaced knots over $\mathbf{x}_1$ and $\mathbf{x}_2$, defined as the Box-Product or "row-wise" Kronecker product of the individual basis (denoted by $\square$):

$$\mathbf{B} = \mathbf{B}_2 \square \mathbf{B}_1 = (\mathbf{B}_2 \otimes \mathbf{1}'_{c_1}) \odot (\mathbf{1}'_{c_2} \otimes \mathbf{B}_1), \tag{5}$$

   where $\mathbf{B}_1$ and $\mathbf{B}_2$ are the B-spline basis along $(\mathbf{x}_1)$ and $(\mathbf{x}_2)$ coordinates.

2. The bidimensional penalty matrix is based on the penalty matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ associated with each marginal basis:

$$\mathbf{P} = \mathbf{P}_2 \otimes \mathbf{I}_{c_1} + \mathbf{I}_{c_2} \otimes \mathbf{P}_1, \tag{6}$$

   where $\mathbf{P}_1 = \lambda_1 \mathbf{D}_1' \mathbf{D}_1$ and $\mathbf{P}_2 = \lambda_2 \mathbf{D}_2' \mathbf{D}_2$, λ_1 and λ_2 being smoothing

parameters which tune the smoothness in each direction, allowing for *anisotropy*.

3. The mixed model matrices turn into:

$$\mathbf{X} = (\mathbf{1}_n : \mathbf{1}_n \square\ \mathbf{x}_1 : \mathbf{x}_2 \square\ \mathbf{1}_n : \mathbf{x}_2 \square\ \mathbf{x}_1) \tag{7}$$

$$\mathbf{Z} = (\mathbf{Z}_2 \square \mathbf{1}_n : \mathbf{Z}_2 \square\ \mathbf{x}_1 : \mathbf{1}_n \square\ \mathbf{Z}_1 : \mathbf{x}_2 \square\ \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1)\,, \tag{8}$$

where matrices $\mathbf{Z}_i$, $i = 1, 2$ are the matrices for random effects (defined in (2)) associated with each of the marginal basis.

The covariance matrix of the random effects becomes now:

$$\mathbf{G} = \begin{pmatrix} \lambda_2 \boldsymbol{\Sigma}_{s,2} \otimes \mathbf{I}_2 & & \\ & \lambda_1 \mathbf{I}_2 \otimes \boldsymbol{\Sigma}_{s,1} & \\ & & \lambda_1 \mathbf{I}_{c_2-2} \otimes \boldsymbol{\Sigma}_{s,1} + \lambda_2 \boldsymbol{\Sigma}_{s,2} \otimes \mathbf{I}_{c_1-2} \end{pmatrix}^{-1}. \tag{9}$$

3. The P-spline family of spatial regression models

In this Section we present the P-spline family of spatial regression models. That is, we extend model (4) by including a short-range spatial dependence term or the traditional econometric specifications by including a non-parametric spatial trend. This way the family we propose accounts for both large and short scale spatial dependence. This new family is defined by the equation:

$$\begin{aligned} \mathbf{y} &= \mathbf{B}\boldsymbol{\theta} + \rho \mathbf{W}_n \mathbf{y} + \mathbf{u} \\ \mathbf{u} &= \delta \mathbf{W}_n \mathbf{u} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \end{aligned} \tag{10}$$

where $\mathbf{B} \equiv \mathbf{B}(\mathbf{x}_1, \mathbf{x}_2)$ is the bidimensional regression basis[1], $\mathbf{W}_n$ is a row-standardized spatial weights matrix which takes into account the n closest neighbours, $\mathbf{W}_n\mathbf{y}$ captures the spatial lag of the dependent variable, ρ is a spatial parameter that measures the existing spatial dependence of $\mathbf{y}$, and δ is a spatial parameter accounting for the spatial dependence in the noise term. In this paper we do not allow that both parameters (ρ and δ) are simultaneously different from zero because this would lead to difficulties in identifying them and, as a consequence, serious problems in the estimation process. Finally, $\boldsymbol{\theta}$ has length $c_1 \mathrm{x} c_2$, which indicates that the model is clearly over-parameterized. This is the reason why the coefficients are penalized according to the expression $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$, with $\mathbf{P}$ defined in Eq. (6).

The model (10) can also be presented as a linear mixed model:

$$\mathbf{A}_1\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad \boldsymbol{\alpha} \sim N(0, \mathbf{G}) \quad \mathbf{u} \sim N(0, \sigma^2 \mathbf{A}_2^{-1}(\mathbf{A}_2^{-1})'). \tag{11}$$

Depending on the values of the spatial parameters, the general spatial regression model encompasses many other models as particular cases.

1. $\rho = 0, \delta = 0 \Longrightarrow$ pure P-spline model (PS) (see Eilers and Marx, 1996).
2. $\rho \neq 0, \delta = 0 \Longrightarrow$ PS-SAR model (see Montero et al, 2012).
3. $\rho = 0, \delta \neq 0 \Longrightarrow$ PS-SEM model (see Lee and Durbán, 2009, where this model is named Smooth-CAR model).
4. $\lambda_1 \to \infty, \lambda_2 \to \infty$ (these parameters are included in matrix $\mathbf{G}$) $\Longrightarrow$ spatial regression models (OLS, SAR or SEM) augmented with a linear global trend over the spatial coordinates.

[1] Of course this term can also be extended to include the effect of other covariates, apart from spatial coordinates, either in a non-parametric or parametric way.

For each type of model the $\mathbf{A}_i$ matrices become:

$$\mathbf{A}_1 = \begin{cases} \mathbf{I}_n - \rho \mathbf{W}_n & \text{in PS-SAR} \\ \mathbf{I}_n & \text{in PS and PS-SEM} \end{cases}$$

$$\mathbf{A}_2 = \begin{cases} \mathbf{I}_n - \delta \mathbf{W}_n & \text{in PS-SEM} \\ \mathbf{I}_n & \text{in PS and PS-SAR} \end{cases}$$

and the other matrices involved in (11) are defined in equations (7) to (9).

As seen in point 4, it is of note that the traditional spatial econometric specifications are also included in the general model when the smoothing spatial parameters λ_1, λ_2 go to infinity. Therefore, specification (10) should be able to have a good performance when the spatial econometric models are suitable (usually when there is no spatial trend or, at most, this spatial trend is linear) and, moreover, improve the performance of the traditional spatial econometric models when non-linearities are present in the spatial trend. Nevertheless, when the spatial trend is close to the linear case, it is very frequent to have some numerical problems in the estimation of the smoothing parameters λ_1 and λ_2.

The parameter vector $(\lambda_1 = \frac{\sigma^2}{\sigma^2_{\alpha_1}}, \lambda_2 = \frac{\sigma^2}{\sigma^2_{\alpha_2}}, \sigma^2, \rho, \delta)$ can be estimated by modifying the REML function as follows:

$$\begin{aligned} \ell_R(\lambda_1, \lambda_2, \sigma^2, \rho, \delta) = &- \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \log |\mathbf{A}_1| \\ &- \frac{1}{2} \mathbf{y}' \mathbf{A}_1' (\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}) \mathbf{A}_1 \mathbf{y}, \end{aligned} \tag{12}$$

where

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma^2 \mathbf{A}_2^{-1} (\mathbf{A}_2^{-1})'. \tag{13}$$

The modified REML function above can be maximized using numerical algorithms. As usual, the vector of fixed effects, $\boldsymbol{\beta}$, and the vector of random

effects, $\boldsymbol{\alpha}$, are estimated as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{A}}_1\mathbf{y} \quad (14)$$

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\hat{\mathbf{A}}_1\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (15)$$

4. Simulation study

In this Section, the performance of all type of P-spline models (PS, PS-SAR and PS-SEM) is compared to the performance of the traditional spatial econometric specifications (SAR and SEM) when the data generating process (DGP) includes a non-linear and non-separable trend (the usual real case), and also when it includes a linear trend. For this purpose, we first simulated 3600 datasets generated by a data generating process including a non-linear and non-separable spatial trend, a spatial lag of the dependent variable and spatial dependence in the noise. Then, for comparison purposes, we repeated this simulating process for the linear trend case. The values of the spatial parameters ρ and δ were chosen to include all the possibilities of P-spline models (PS, PS-SAR and PS-SEM).

More specifically, the data generating process[2] (DGP) is given by:

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}_1, \mathbf{x}_2) + \rho\mathbf{W}_n\mathbf{y} + \mathbf{u} \\ \mathbf{u} &= \delta\mathbf{W}_n\mathbf{u} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \end{aligned} \quad (16)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ were generated by a uniform distribution in $(0, 1)$. When the trend is non-linear, $f(\mathbf{x}_1, \mathbf{x}_2)$ is specified as in Wood (2006), that is:

$$\begin{aligned} f(\mathbf{x}_1, \mathbf{x}_2) = 10\pi\sigma_{x_1}\sigma_{x_2} \{1.2\exp\left(-(\mathbf{x}_1 - 0.2)^2/\sigma_{x_1}^2 - (\mathbf{x}_2 - 0.3)^2/\sigma_{x_2}^2\right) + \\ +0.8\exp\left(-(\mathbf{x}_1 - 0.7)^2/\sigma_{x_1}^2 - (\mathbf{x}_2 - 0.8)^2/\sigma_{x_2}^2\right)\}, \end{aligned} \quad (17)$$

[2] All computations were made using R software. More specifically, we used the packages spline and spdep (see R Development Core Team, 2013; Bivand, 2013, respectively). The codes are available upon request.

with $\sigma_{x_1} = 0.3$ and $\sigma_{x_2} = 0.4$.

The parameter values chosen for the simulation procedure are $\rho = (0, 0.5, 0.75)$, $\delta = (0, 0.5)$, $\sigma = 0.5$, and the true number of neighbours in the matrix $\mathbf{W}$ is 6. This way, we encompass all the P-spline models in the DGP: PS when $\rho = 0, \delta = 0$; PS-SAR when $\rho \neq 0, \delta = 0$ and PS-SEM when $\rho = 0, \delta \neq 0$. There are also some DGP's not corresponding with any considered P-spline model when $\rho \neq 0, \delta \neq 0$. This way we can evaluate the performance of each model for particular and general cases. Furthermore, to evaluate the effects of under-specification or over-specification of the number of neighbours in the matrix $\mathbf{W}$, we estimate the whole set of models with 3, 6 and 10 neighbours. The simulation algorithm is composed of the following five steps:

1. Generate the random vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ (length $n = 200$). Compute the trend values, which remain fixed in all the simulations. Generate the contiguity matrix $\mathbf{W}_n$ on the basis of the simulated coordinates. The elements of $\mathbf{W}_n$ are 1 for the six closest neighbours and 0 otherwise. As usual, $\mathbf{W}_n$ is used in row-standardized form.

2. Choose a couple of values for both ρ and δ and generate n values of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Choose a number of neighbours (3, 6 or 10) to account for the consequences of under-specification, correct specification or over-specification of the number of neighbours. Then, compute the values of vectors $\mathbf{u}$ and $\mathbf{y}$ as in Eq. (16).

3. Make a REML estimation of the model parameters for all P-spline models considered (PS, PS-SAR and PS-SEM). The matrix of regressors $\mathbf{B}(\mathbf{x}_1, \mathbf{x}_2)$ is built with B-spline basis matrices, the number of knots being set at 10. As usual in this type of models, second-order penalties are used. Make also a ML estimation of the spatial econometric models (SAR and SEM) including $\mathbf{x}_1$ and $\mathbf{x}_2$ as independent regressors (more specifically, as if a linear trend was known). The contiguity matrix includes the number of neighbours specified in step (2).

4. Return to step (2) and repeat the process $m = 200$ times for each possible combination of ρ, δ and the number of neighbours.

5. Compute, as follows, the mean squared error (MSE) for both the estimated trend values and the estimated values of the response variable.

$$\text{MSE}_{\text{trend}} = \sum \left[f(\mathbf{x}_1, \mathbf{x}_2) - \widehat{f(\mathbf{x}_1, \mathbf{x}_2)} \right]^2 / n, \tag{18}$$

where

$$\widehat{f(\mathbf{x}_1, \mathbf{x}_2)} = \begin{cases} \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha} \text{ in PS, PS-SAR and PS-SEM} \\ \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + \hat{\beta}_2 \mathbf{x}_2 \text{ in SAR and SEM,} \end{cases}$$

and

$$\text{MSE}_{\text{y-obs.}} = \sum \left[\left(\hat{\mathbf{A}}_1 \mathbf{y} - \widehat{f(\mathbf{x}_1, \mathbf{x}_2)} \right)' \left(\hat{\mathbf{A}}_2' \hat{\mathbf{A}}_2 \right) \left(\hat{\mathbf{A}}_1 \mathbf{y} - \widehat{f(\mathbf{x}_1, \mathbf{x}_2)} \right) \right]^2 / n, \tag{19}$$

where

$$\hat{\mathbf{A}}_1 = \begin{cases} \mathbf{I}_n \text{ in PS, PS-SEM and SEM} \\ \mathbf{I}_n - \hat{\rho}\mathbf{W}_n \text{ in PS-SAR and SAR,} \end{cases}$$

and

$$\hat{\mathbf{A}}_2 = \begin{cases} \mathbf{I}_n \text{ in PS, PS-SAR and SAR} \\ \mathbf{I}_n - \hat{\delta}\mathbf{W}_n \text{ in PS-SEM and SEM.} \end{cases}$$

Figures 1 to 4 summarize the simulation results. In Figures 3 and 4 we use the logarithmic scale for the sake of better visualization. Next, we highlight the main findings regarding the spatial parameters, the specified number of neighbours, the estimation of the trend and the estimation of the observed

(simulated) values of the dependent variable.

Figure 1: Estimates of ρ in PS-SAR and SAR models in the simulation process (non-linear trend)

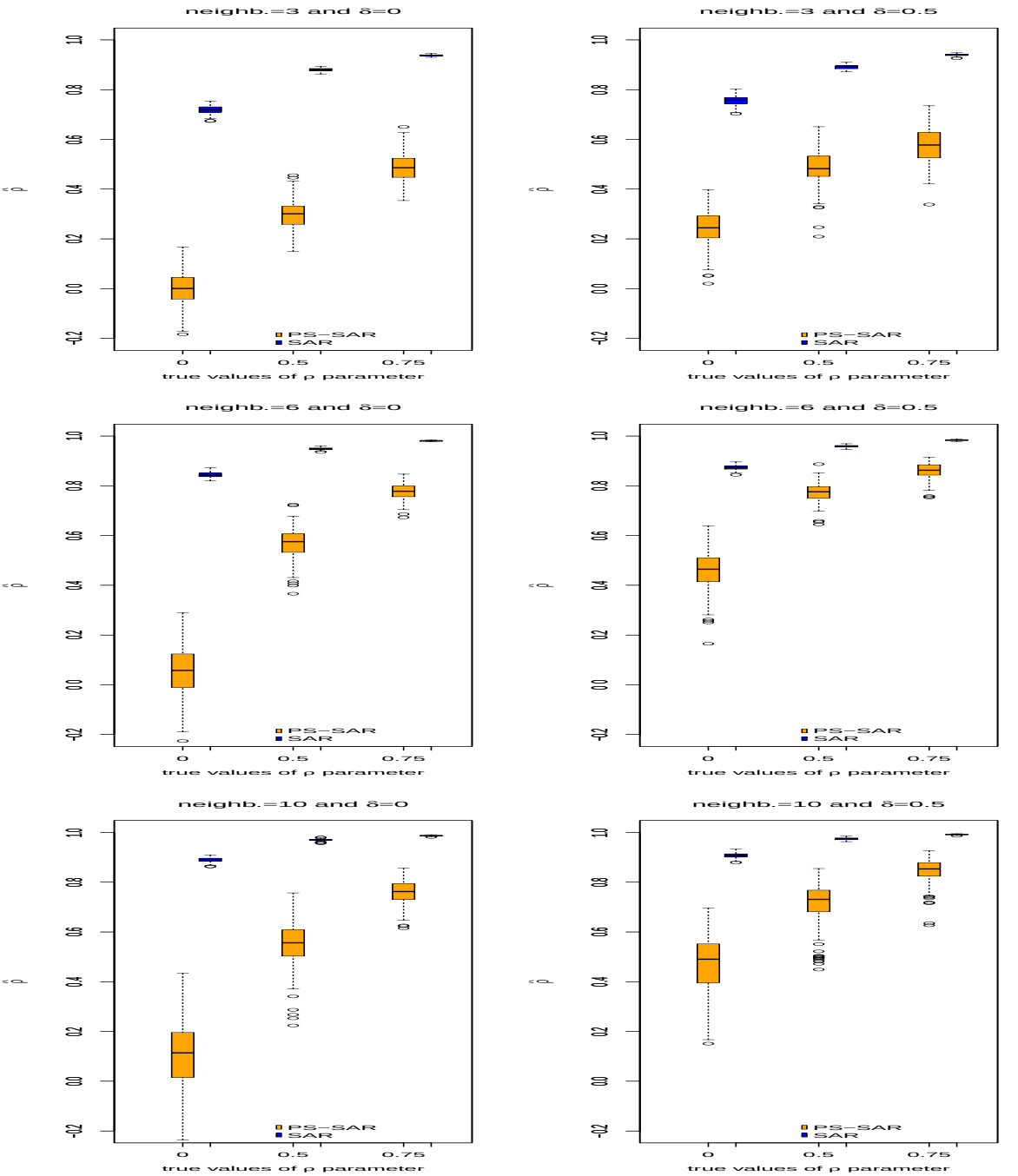


Figure 2: Estimates of δ in PS-SEM and SEM models in the simulation process (non-linear trend)

Figure 3: LogMSE in the estimation of **y**-observed by simulation (non-linear trend)

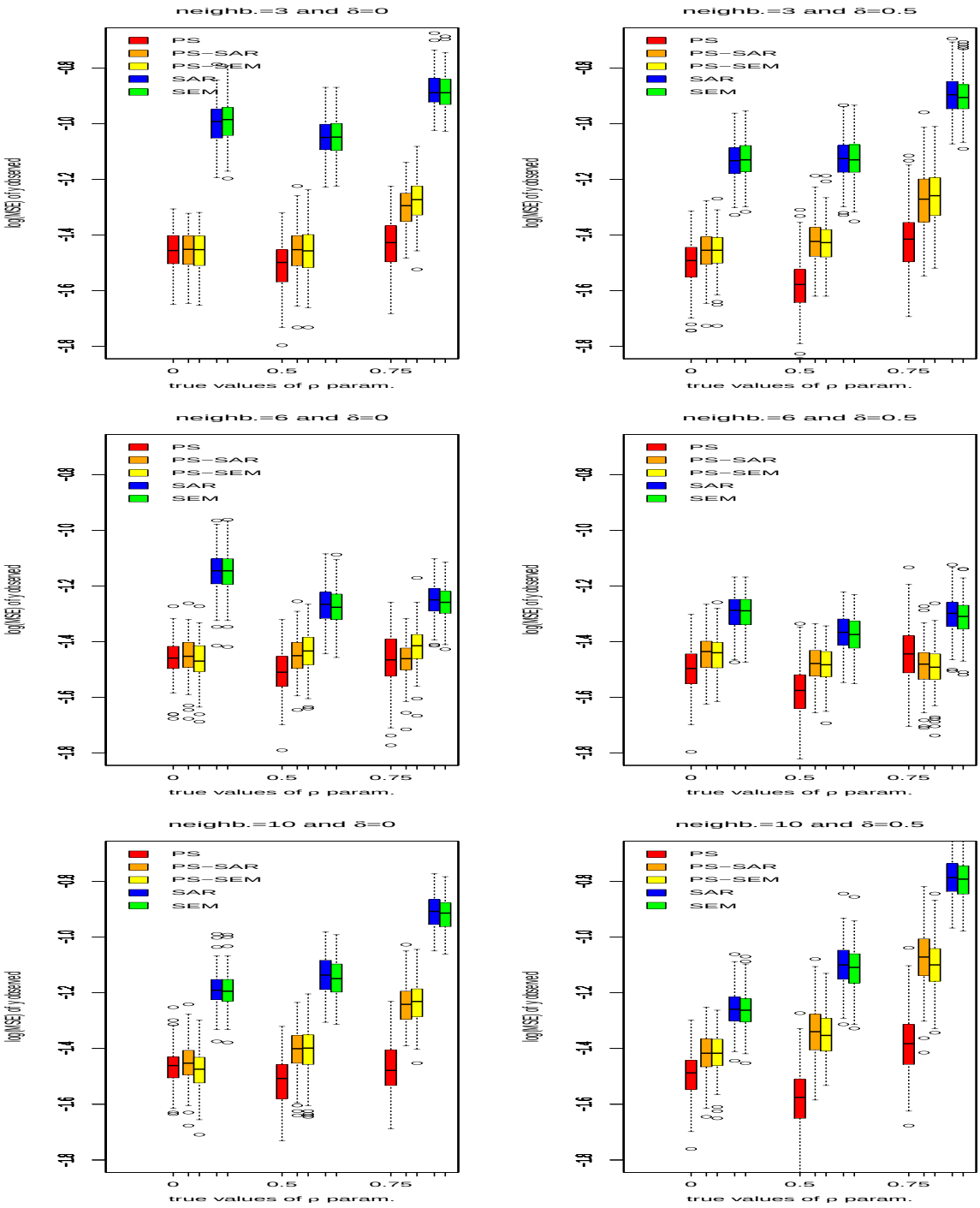
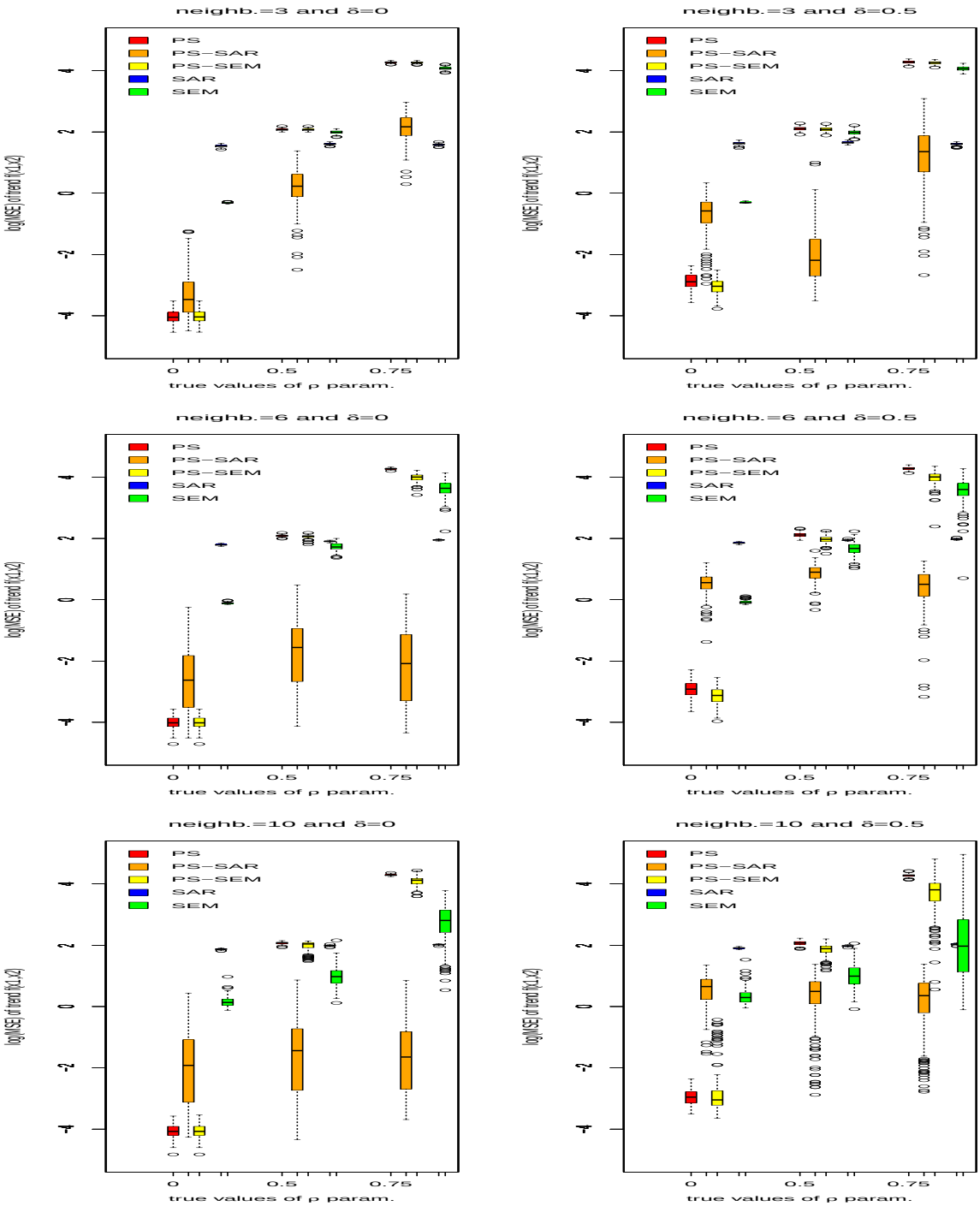


Figure 4: Log-MSE in the estimation of the trend by simulation (non-linear trend)



The main conclusion extracted from them are the following:.

1. The spatial parameters ρ and δ are much better estimated for PS-SAR and PS-SEM than for SAR and SEM models. Logically, the linear specifications overestimate the spatial parameters due to the misspecification of the spatial trend. Comparing the P-spline specifications, the estimates of ρ in PS-SAR have less dispersion than the estimates of δ in PS-SEM. This could be due to the greater difficulties of identification of the spatial correlation in the error term. In general the under-specification of the number of neighbours has worse consequences than the over-specification of it.

2. The log-MSE in the estimation of the $\mathbf{y}$-observed values is lower for P-spline type models. It is remarkable the better performance of the pure PS model in case of over-specification of the number of neighbours.

3. Finally, when analyzing the log-MSE in the estimation of the trend there is, as expected, a huge difference between the P-spline type models and the classical spatial regression models in favour of the P-spline family of models. Obviously, the pure PS model only has a good performance when $\rho = 0$.

As stated at the beginning of the section, for the purpose of evaluating the performance of P-spline models when the data generating process includes a linear trend instead of a non-linear spatial one, we have repeated the simulation exercise including the following linear trend:

$$f(\mathbf{x}_1, \mathbf{x}_2) = 2.5 - 0.5\mathbf{x}_1 + 0.1\mathbf{x}_2 \tag{20}$$

This spatial trend has been embedded in (16) to simulate new data and reestimate the whole set of models. The corresponding figures are available under

request as additional material. For the linear trend case, the following conclusions can be extracted:

1. There are no differences in the estimates of spatial parameters for both types of specifications (P-spline and linear econometric models). This shows the robustness of PS-SAR and PS-SEM estimates also when the true spatial trend is linear.

2. Furthermore, the log-MSE in the estimation of the **y**-observed values is similar in both linear and P-spline models. Quite surprisingly, the performance of PS model is especially good, particularly for high values of ρ and δ.

3. Finally, the estimation of the (linear) trend is similar in both traditional econometric specifications and P-spline models. Consequently, the performance of both tradional and new models only depends on whether they include or not the ρ, δ, or both parameters.

To sum up, the P-spline specifications (PS, PS-SAR and PS-SEM) have a better behaviour when the true spatial trend is non-linear and, moreover, when the true spatial trend is linear, their performance is similar to the linear econometric specifications (SAR and SEM). Nevertheless, PS model is unable to identify the true trend when either ρ or δ are different from zero. In practical terms, the non-parametric models seem to have a better behaviour than traditional specifications and should be seriously considered as good alternatives to them.

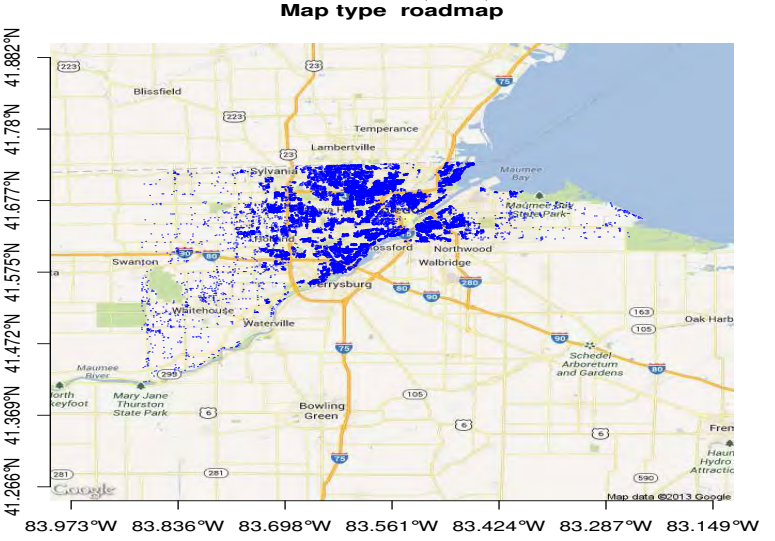
5. Empirical Case

In this section the performance of the pure PS, PS-SAR and PS-SEM models is studied by using the well-known Lucas County (Ohio) database on house prices. This way our results can be compared with previous research based on traditional spatial strategies. Obviously, since now we use real data instead of a

data generating process, the trend is unknown and the performance of the proposed P-spline models must be analyzed in terms of the MSE and Information Criteria when predicting the observed variable.

Applications of spatial econometrics to real estate valuation are certainly scarce, and as a consequence, the Lucas County database on house sales has not been extensively used in the literature on the topic. However, it constitutes the database used in some of the most prominent research applying spatial econometric models, see LeSage and Pace (2009); Bivand (2010); Zhu et al (2011); Dubé and Legros (2013).

Figure 5: Lucas County Map (Ohio) including houses observations.



Source: GoogleMap and spdep package (Bivand, 2013)

The Lucas County, Ohio, housing data set has 18,378 observations of single family homes sold during 1995-1998, and is fully described in the file data/house.txt in the Spatial Econometrics toolbox (see Figure 5). The dependent variable is the logarithm of the selling price and, for parametric specifications, the regressors include the age, the squared age, the logarithms of the lot size and the total living area, and the number of rooms and bedrooms. The cubed age of the house is usually included in the set of independent variables, but we did

not consider it because of its non significance. As stated in Bivand (2012), no contextual variables about the neighborhood of the houses are available, so one would expect a strong spatial autocorrelation reflecting this misspecification.

The list of neighbours provided with the data set in *spdep* is a sphere of influence graph constructed from a triangulation of the point coordinates of the houses after projection to the Ohio North NAD83 (HARN) Lambert Conformal Conical specification (EPSG:2834). It is relatively sparse, with less than three neighbours per observation on average.

Finally, for P-spline models, the spatial trend, $f(x_1, x_2)$, is constructed by using x_1 and x_2 coordinates in rescaled form. The rest of regressors are also included in a non-parametric way.

The Lucas County house data set has been widely used for different purposes. Here we use it to compare the performance, in terms of MSE and Information Criteria, of 6 competing models: 1) The regression model (OLS); 2) SAR model; 3) SEM model; 4) pure PS model; 5) PS-SAR model and 6) PS-SEM model. The three first models (OLS, SAR and SEM) include a spatial linear trend for comparison purposes.

As can be seen in Table 1, the estimates of the spatial autoregressive parameters ρ and δ are certainly large. More specifically, ρ ranges between 0.35 and 0.40 in PS-SAR and between 0.46 and 0.51 in SAR. Something similar happens with δ: it ranges from 0.42 to 0.48 when a PS-SEM is estimated, and from 0.52 to 0.64 when the estimated strategy is a SEM. These results give an idea of the importance of including a non-parametric trend term in the model. Otherwise, the spatial relations due to the existence of a complex spatial trend are attributed to the spatial corresponding autoregressive parameter, making it greater. We have included a linear trend in SAR and SEM models but the shape of the true trend is unknown; this drawback can be easily overcome by using a P-spline approach, which does not require a pre-specified shape of the trend, but captures it from the data.

Table 1: Spatial parameters estimated for each type of model.

		$\hat{\rho}$		$\hat{\delta}$	
Year	n	SAR	PS-SAR	SEM	PS-SEM
1995	4130	0.50	0.38	0.62	0.44
1996	4838	0.46	0.34	0.52	0.44
1997	5032	0.46	0.35	0.54	0.42
1998	4378	0.51	0.40	0.64	0.48

Table 2 lists the MSE committed when estimating the observed values with each of the above referred specifications. We have also computed the Bayesian Information Criterion (BIC) to penalize the possible overfitting, due to the increase of degrees of freedom, of P-spline compared with traditional specifications. Clearly the best options, both in terms of MSE and BIC, are PS-SAR and PS-SEM models. They improve the performance, not only of SAR and SEM, but also of pure PS specification. For comparison purposes, we have also estimated the additive specifications corresponding to P-spline type models. Additive models are particular cases of P-spline ones which do not allow interaction between the coordinates in the spatial trend. Nevertheless, the additive specifications never improve the general P-spline models which reinforces the thesis that these interactions are really important and the true spatial trend could be non-linear and non-separable.

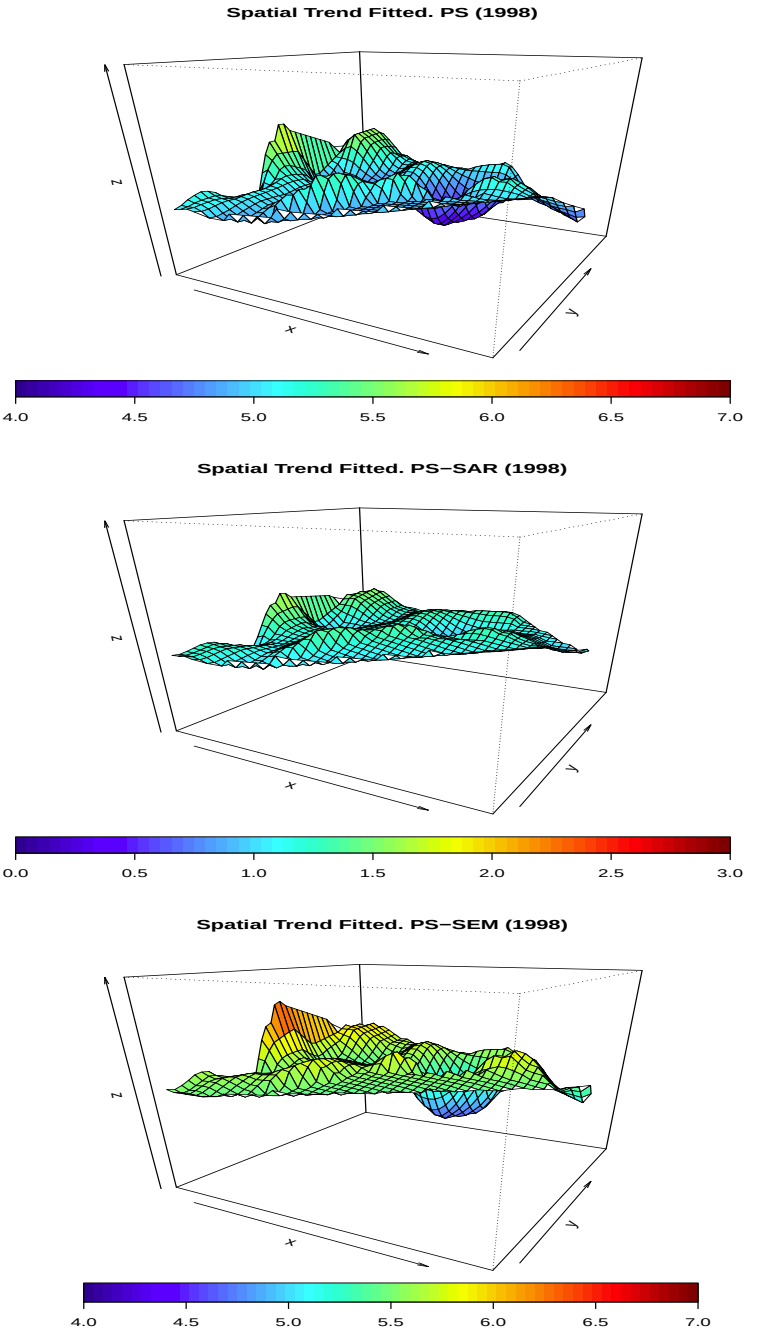
Table 2: Goodness-of-Fit and Information Criteria of estimated models for Lucas County (Ohio), 1995-1998.

				Mean Squared Errors (MSE)					
				Non-Additive Models			**Additive Models**		
Year	OLS	SAR	SEM	PS	PS-SAR	PS-SEM	PS	PS-SAR	PS-SEM
1995	0.157	0.085	0.088	0.094	**0.070**	0.074	0.111	0.074	0.078
1996	0.214	0.138	0.147	0.127	0.103	**0.100**	0.145	0.108	0.105
1997	0.196	0.124	0.126	0.124	**0.099**	**0.099**	0.140	0.103	0.103
1998	0.169	0.089	0.091	0.107	**0.078**	0.080	0.128	0.082	0.085
Bayesian Information Criteria (BIC)									
1995	6.489	5.875	5.911	6.065	**5.758**	5.812	6.190	5.773	5.833
1996	6.958	6.517	6.578	6.521	6.296	**6.270**	6.604	6.306	6.278
1997	6.905	6.447	6.467	6.521	6.278	**6.276**	6.607	6.291	6.293
1998	6.624	5.982	5.998	6.251	**5.907**	5.941	6.388	5.925	5.959

Although there are no major differences between the goodness-of-fit of PS-

SAR and PS-SEM models, in terms of computation, it is much easier to estimate PS-SAR than PS-SEM model (with 4000-5000 observations the difference can reach some hours of computation time). In practical terms, the best option to specify seems to be the PS-SAR model because it is enough flexible, robust to the existence of complex spatial trends and short-range spatial correlation and is easy to estimate. Figure 6 shows the spatial trend fitted for all P-spline models in 1998.

Figure 6: Spatial trends fitted for P-spline models in 1998.



The scale for PS-SAR model is different than for PS and PS-SEM models.

Nevertheless, the range of scale is the same in all cases for comparative purposes.

In brief, the results obtained for this empirical case reinforce the main conclusions derived from simulations: PS-SAR and PS-SEM models outperform the corresponding traditional spatial specifications in terms of MSE and BIC, do not mix large-scale dependencies with local spatial autocorrelation, and do not need a pre-specified shape of the trend. Moreover, although in the simulation study the performance of pure PS was good enough in terms of MSE, in this empirical case it is clearly outperformed by PS-SAR and PS-SEM specifications. This fact could be explained for the existence of a significant local spatial autocorrelation. In addition, the exclusion of the interaction term from the trend (additive specifications) is not a good choice, since it increases significantly the MSE and BIC. This is due to the complexity of the existing trend (something very common in real phenomena that evolve across the space), which can be only captured by including such interaction term.

6. Conclusions

Most of phenomena that evolve in space, the more and more frequently studied in a large variety of scientific disciplines, usually show a complex non-linear trend which depends on both the geographical coordinates and the interaction between them. However, the traditional spatial econometric models have serious difficulties to incorporate this circumstance. This is why in this article we augment the family of traditional spatial econometric models by including a complex non-linear and non-separable trend term. Our approach is to capture the existing trend from the data in a non-parametric way by using a B-splines regression basis, a penalty on the coefficients and to add a spatial econometric strategy to the trend term. The local spatial autocorrelation is estimated together with the large-scale spatial dependencies and, moreover, the shape of the trend is controlled by two smoothing parameters which allow for anisotropy, making the model even more flexible. We use penalized splines to overcome the over-parameterization problem and represent the augmented model as a mixed

model. The smoothing parameters can be estimated along with the other parameters of the model. We call this new family of models P-spline spatial regression models. This family includes the pure P-spline model and PS-SAR and PS-SEM models.

To evaluate the performance of the new family of spatial econometric models, we firstly proceed to simulate 3600 data sets generated by the foregoing P-splines strategies, where the trend was complex, non-linear and non-separable. We also consider a large variety of combinations of the parameters of the model and different number of neighbors. From this simulation exercise it can be concluded that the new class of spatial econometric models reproduce the spatial autoregressive parameters set in the data generating process much better than the traditional strategies. In addition, the new strategies far outweigh the traditional ones when it comes to estimating the observed (simulated) values and, especially, the complex, non-linear and non-separable trend incorporated in the data generating process.

The performance of the new family of models was also checked by using real data. In particular, we use the well-known Lucas County (Ohio) data set of house prices. The results obtained from this empirical case confirm those obtained in the simulation exercise: the P-spline specifications outperform the corresponding traditional strategies.

The good results obtained by both simulating and studying a real case encourage us to go further. Some interesting avenues of future research we suggest include the analysis of a large and varied number of empirical cases, the interpretation of the spatial parameters and the study of the identification conditions in the PSAC case (the most general P-spline strategy we consider), the extension of our methodology to the spatio-temporal case, and the incorporation of the P-spline methodology to the smoothing-spline-ANOVA models.

References

Bivand R (2010) Comparing estimation methods for spatial econometrics techniques using R, Discussion paper 2010:26, Department of Economics, Norwegian School of Economics and Business Administration

Bivand R (2012) After Raising the Bar: applied maximum likelihood estimation of families of models in spatial econometrics. Estadística Española 54(177):71–88

Bivand R (2013) spdep: Spatial dependence, weighting schemes, statistics and models. URL http://CRAN.R-project.org/package=spdep/, R package version 0.5-61

Cressie N, Wikle CK (2011) Statistics for Spatio-Temporal Data. Wiley

Currie ID, Durbán M (2002) Flexible smoothing with P-splines: A unified approach. Statistical Modelling 2:333–349

De Boor C (1977) Package for calculating with B-splines. Journal of Numerical Analysis 14:441–472

Dubé J, Legros D (2013) Dealing with spatial data pooled over time in statistical models. Letters in Spatial and Resource Sciences 6(1):1–18

Eilers P, Marx B (1996) Flexible smoothing with B-splines and penalties. Statistical Science 11:89–121

Lee DJ, Durbán M (2009) Smooth-car mixed models for spatial count data. Computational Statistics and Data Analysis 53:2968–2977

LeSage J, Pace R (2009) Introduction to Spatial Econometrics. Chapman-Hall, Boca Raton (USA)

Montero J, Mínguez R, Durbán M (2012) SAR models with nonparametric spatial trends. A P-spline approach. Estadística Española 54(177):89–112

R Development Core Team (2013) R: A language and environment for statistical computing. URL http://www.R-project.org/

Wood S (2006) Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. Biometrics 62:1025–1036

Zhu B, Füss R, Rottke NB (2011) The predictive power of anisotropic spatial correlation modeling in housing prices. The Journal of Real Estate Finance and Economics 42:542–565

Modelling long term trend and local spatial correlation: a mixed penalized spline and spatial econometrics approach.

by

Román Mínguez

Universidad de Castilla-La Mancha (Spain)[*]

María Durbán

Universidad Carlos III de Madrid (Spain)[**]

José-María Montero

Universidad de Castilla-La Mancha (Spain)[***]

Dae-Jin Lee

CSIRO-CMIS (Australia)[****]

[*]Corresponding author: Roman.Minguez@uclm.es

[**]maria.durban@est-econ.uc3m.es

[***]Jose.MLorenzo@uclm.es

[****] Dae-Jin.Lee@csiro.au